图优化的低秩双随机分解聚类 *

张 涛，胡恩良†，于景丽

(云南师范大学 数学学院，昆明 650500)

摘 要：低秩双随机矩阵分解聚类(low-rank doubly stochastic matrix decomposition for cluster analysis，DCD)是最近由Yang等人[16]提出的一种图聚类方法，它通过最小化KL(Kullback-Leibler)散度准则：KL(A，S)，从图关联矩阵S中获得一个非负低秩双随机矩阵分解：A=UUT(U≥0)，并以U作为类标签矩阵进行聚类。在DCD方法中，因矩阵S是固定不可变的，故S初始取值选取的好坏对聚类结果有极大影响，这导致了它缺乏稳定性。针对这一问题，提出了一种基于图优化的DCD方法，将图关联矩阵S和DCD的优化集成在统一框架中，这改进和拓展了原始的DCD方法。实验结果表明，与DCD方法相比，图优化的DCD方法具有更好的聚类精确度和稳定性。

关键词：低秩双随机矩阵分解；图优化；稳定性；聚类

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2017.08.0874

Graph-optimized low-rank doubly stochastic decomposition for clustering

Zhang Tao, Hu Enliang†, Yu Jingli

(*Department of mathematics Yunnan Normal University*, *Kunming Yunnan* 650500, *China*)

Abstract: Clustering by DCD (low-rank doubly stochastic matrix decomposition) was recently proposed by Yang[16] as a method of graph clustering. DCD obtains a nonnegative low-rank doubly stochastic decomposition A=UUT(U≥0) from the graph correlation matrix S by minimizing the criterion of KL (Kullback-Leibler) divergence: KL (A, S) , and clustering from U, as the class label matrix. In the method of DCD, because the S is pre-fixed, the initial value of S has a great influence on the clustering result, which leads to its lack of stability. Aiming at this problem, propose a DCD method based on graph optimization , and the optimization of graph correlation matrix S and DCD is integrated in a unified framework, which improves and extends the original DCD. The experimental results show that the graph-optimized DCD has better clustering accuracy and stability than the original DCD.

Key Words: low-rank doubly stochastic matrix; graph optimization; stability; clustering

0 引言

聚类是根据“物以类聚”思想,将本身没有类别的对象聚集成不同的簇,并且对每一个这样的簇进行描述的过程。聚类的目的是使得属于同一个簇的对象之间彼此相似,而不同簇之间的对象足够不相似。聚类分析是机器学习、数据挖掘和模式识别等领域的重要研究内容之一。根据方法类型,聚类算法大体可以分为以下几类：基于划分的方法,如K-means[1]、K-medoids[2]等；基于层次的方法,如CURE[3]等；基于网格的方法,如STING[4]等；基于密度的方法,如DBSCAN[5]等；基于神经网络的方法,如SOM[6]等；基于图的方法,如Normalized cut[7]等。不同聚类方法拥有各自的优点,但在一定程度上也都存在各自的缺点,因此探索新的聚类方法具有重要意义。本文提出的新聚类方法属于基于图的聚类方法。

1 相关背景介绍

1.1 图聚类

图聚类算法[8,9]是建立在图理论基础上,其本质是先用图来表示对象之间的关系,再将聚类问题转换为图划分问题,这是一种点对聚类算法。在图聚类中,对象间的图结构由一个关联矩阵来表达,图构建的质量将最终决定聚类结果的好坏。图构建过程通常包括图的边选择与边权配置两步。广泛使用的边构造方式有K近邻图[10]、ε球近邻图[11]和全连接图等。图的边建成后,边权配置[12]方式也多种多样,其中使用较多的方法是0-1二值权重和利用热核函数的权重设置[13]等。

1.2 低秩双随机矩阵分解聚类

在过去的10年里,低秩矩阵分解技术逐渐在机器学习与数据挖掘领域获得诸多应用。特别地,非负低秩矩阵分解技术已成

基金项目：国家自然科学基金资助项目（61663049，61165012）；云南师范大学研究生科研创新基金项目（yjs201678）

作者简介：张涛（1992-），男，安徽六安人，硕士研究生，主要研究方向为机器学习及数据挖掘；胡恩良（1975-），男（通信作者），副教授，博士，主要研究方向为机器学习和最优化（humath@ynnu.edu.cn）；余景丽（1990-），女，硕士研究生，主要研究方向为机器学习及数据挖掘.

功应用于聚类方面。1999 年,Huffman[14]提出利用概率潜语意指示来分割数据,矩阵分解中使用 KL(Kullback-Leibler)散度代替传统的欧氏距离。2001 年,Lee 等人[15]提出的非负矩阵分解方法将矩阵成对分解成 2 个非负低秩矩阵的乘积形式。2010 年，Ding 等人[19]提出非负矩阵分解近似传统 K-means 方法。2013 年，Arorat 等人[20]提出左随机矩阵分解近似于左随机矩阵所产生的相似矩阵。最近,由 Yang 等人[16]提出了一种非负低秩双随机矩阵分解(low-rank doubly stochastic matrix decomposition,DCD)的图聚类方法,DCD 的主要思想是：最小化图关联矩阵和一个低秩双随机矩阵之间的 KL 散度,其中双随机矩阵由聚类标签矩阵的乘积构成。若记 rank(U)=r,则 r-秩双随机矩阵集合可表示如下：

$$\mathbb{A}=\left\{A \mid A=UU^{T}, U\geq 0, \sum_{j} A_{ij}=1\right\}.$$

若记 rank(W)=r 且

$$\mathbb{B}=\left\{B \mid B_{ij}=\sum_{k=1}^{r}\frac{W_{ik}W_{jk}}{\sum_{v}W_{vk}}, \sum_{k=1}^{r}W_{ik}=1, W\geq 0\right\}$$

,Yang 等人[16]证明了以上集合 $\mathbb{A}$ 与 $\mathbb{B}$ 等价,即有如下定理。

定理 1[16] $\mathbb{A}\equiv\mathbb{B}$.

以上定理说明,$\mathbb{B}$ 也是双随机矩阵集合。相对于集合 $\mathbb{A}$ 的表示,集合 $\mathbb{B}$ 的表示形式更有利于优化求解,因此本文以下对双随机矩阵集合的表述将基于集合 $\mathbb{B}$。

若设 $W=[w_1,w_2,\cdots w_n]^T\in R^{n\times r}$ 为待求聚类标签矩阵(其中 $w_i\in R^{r\times 1}$), S^0 为初始设定的图关联矩阵,则 DCD 方法对应的优化问题为

$$\min: J(W)=\mathrm{KL}\left(B,S^{0}\right)-(\alpha-1)\log W_{ik} \tag{1}$$

其中：

$$\mathrm{KL}\left(B,S^{0}\right)=\sum_{ij}\left(S_{ij}^{0}\log\frac{S_{ij}^{0}}{B_{ij}}-S_{ij}^{0}+B_{ij}\right),$$

$$B_{ij}=\sum_{k=1}^{r}\frac{W_{ik}W_{jk}}{\sum_{v}W_{vk}}, \sum_{k=1}^{r}W_{ik}=1, W\geq 0\text{，即 } B\in\mathbb{B}.$$

在式(1)中,目标函数 $J(W)$ 的第一项旨在最小化相似矩阵 S^0 与双随机矩阵 B 的 KL 散度,第二项则强化 W 的非负性。为了求解(1),文献[16]中先利用拉格朗日乘子法消除约束,再令新目标函数关于 W 的导数为零,最后利用乘性更新(multiplication update)算法迭代求解聚类标签矩阵 W,具体推导过程请详见文献[16]。

1.3 DCD 聚类方法存在的不足及改进

在图聚类方法中,图的构造是无监督的,因此带有一定的随机性,这将导致以下一些不足：a)图或其对应的关联矩阵 S^0 是人为预先定义的,在后续学习过程中不能被优化;b)图构造时仅利用了原始数据的空间结构,而这种原始结构不一定最有利于后续的聚类任务;c)图构建时涉及到边权重的配置方式,这常导致参数选择困难(例如,利用热核权重时需进行核参数的选择)。

为了解决上述这些不足,受图优化降维研究[17,18]的启发,本文将在第 3 部分提出图优化的低秩双随机矩阵分解聚类(简记为 GoDCD),该方法将图的优化过程合并到DCD 目标函数的优化中,从而获得图(关联矩阵)优化和双随机矩阵分解的同步学习框架。本文的算法优点是:在 GoDCD 中,图构建不是初始固定的,而是随着算法迭代会被逐步优化,因此 GoDCD 能减轻对初始关联矩阵的依赖,寻找到更合适于后续聚类任务的图关联矩阵。

2 图优化的双随机分解聚类

2.1 模型建立

在 DCD 模型中,图构建等价于构造初始的图关联矩阵 S^0。若 S^0 构造得不好,则后续聚类效果会很差。为了部分克服此问题,本文提出的 GoDCD 模型将图优化与 DCD 聚类模型集成到统一的学习框架下,其目标函数为

$$\min: J(W,S)=\mathrm{KL}\left(B,S\right)-(\alpha-1)\log W_{ik}+\lambda\mathrm{KL}\left(S,S^{0}\right) \tag{2}$$

其中:

$$\mathrm{KL}\left(B,S\right)=\sum_{ij}\left(S_{ij}\log\frac{S_{ij}}{B_{ij}}-S_{ij}+B_{ij}\right)$$

$$\mathrm{KL}\left(S,S^{0}\right)=\sum_{ij}\left(S_{ij}^{0}\log\frac{S_{ij}^{0}}{S_{ij}}-S_{ij}^{0}+S_{ij}\right)$$

$$B\in\mathbb{B}.$$

对比问题式(1)(2)中的两个模型,容易看出GoDCD 与DCD 的目标函数区别在于：

a)GoDCD 比 DCD 多了一项,即 $\mathrm{KL}(S,S^0)$,其作用是将在 S^0 的邻域内优化一个比 S^0 更优的关联矩阵 S;

b)对应于 DCD 中的项 $\mathrm{KL}(B,S^0)$ 在 GoDCD 中被替换为 $\mathrm{KL}(B,S)$,其目的是在更优关联矩阵 S(而不是 S^0)的基础上来进行低秩双随机分解聚类。

c)在 DCD 中仅 W 被优化,而在 GoDCD 中 W 和 S 同时被优化,这相当于将图优化和低秩双随机分解集成在同一个目标函数中,其目的是使图构建(对应 S)和聚类(对应 W)达到联合最优。

2.2 模型求解

因为目标函数 $J(W,S)$ 为非凸函数,故求解问题式(2)属于非凸优化问题。对此问题,本文采用交替最小化方法对其迭代求解,即先固定 S,求解关于 W 的子问题;再固定 W,求解关于 S 的子问题,具体如下：

$$W^{(t)}=\arg\min_{W}: J(W,S^{(t-1)}) \tag{3}$$

$$S^{(t)}=\arg\min_{S}: J(W^{(t)},S) \tag{4}$$

由此产生的迭代序列如下：

$$S^{0}\rightarrow W^{(1)}\rightarrow S^{(1)}\rightarrow W^{(2)}\rightarrow S^{(2)}\cdots\rightarrow W^{(t)}\rightarrow S^{(t)}\cdots\text{。}$$

对子问题式(3)的求解,可直接使用文献[16]中的 DCD 求解算法。对于子问题式(4),其解具有封闭形式,具体推导如下：

令 $J=J(W^{(t)},S)$,对 S_{ij} 求偏导数,得

$\frac{\partial J}{\partial S_{ij}}=\log S_{ij}-\log B_{ij}^{(t)}+\lambda(\log S_{ij}-\log S_{ij}^{0})$ 其中(由定理 1),

$\mathrm{B}_{ij}^{(t)}=\sum_{k=1}^{r}\frac{w_{ik}^{(t)}w_{jk}^{(t)}}{\sum_{v=1}^{N}w_{vk}^{(t)}}$。

令 $\frac{\partial J}{\partial S_{ij}}=0$,得

$$(\lambda+1)\log S_{ij}=\log B_{ij}^{(t)}+\lambda\log S_{ij}^{0}$$
$$\Leftrightarrow(\lambda+1)\log S_{ij}=\log B_{ij}^{(t)}+\log(S_{ij}^{0})^{\lambda}$$
$$\Leftrightarrow S_{ij}=\sqrt[\lambda+1]{B_{ij}^{(t)}\times(S_{ij}^{0})^{\lambda}}$$

即得

$$S_{ij}^{(t)}=\left(B_{ij}^{(t)}\times(S_{ij}^{0})^{\lambda}\right)^{\frac{1}{\lambda+1}}.$$

综述所述,对 GoDCD 模型的求解算法可整理为算法 1。

算法 1 *GoDCD* 求解算法

初始化：r ---类别数，$S^{(0)}=S^{0}$，$t=1$，λ.

输出：聚类标签矩阵 W.

Repeat

Step1(W 更新)：利用 DCD 求解算法解子问题：$W^{(t)}=\arg\min_{W}:J(W,S^{(t-1)})$；

Step2(S 更新)：$S_{ij}^{(t)}=\left(B_{ij}^{(t)}\times(S_{ij}^{0})^{\lambda}\right)^{\frac{1}{\lambda+1}}$；

Step3: $t=t+1$;

Until $J(W^{(t-1)},S^{(t-1)})-J(W^{(t)},S^{(t)})\le\varepsilon$

或者 $t>itermax$

输出：$W=W^{(t)}$,结束。

以下定理 2 表明算法 1 是收敛的。

定理 2 若 $\left\{J\left(W^{(t)},S^{(t)}\right)\right\}$ 是由算法 1 产生的序列,则该序列收敛。

证明 由问题式(3)和(4)可知,

$J\left(W^{(t)},S^{(t)}\right)\le J\left(W^{(t)},S^{(t-1)}\right)\le J\left(W^{(t-1)},S^{(t-1)}\right)$，即 $\left\{J\left(W^{(t)},S^{(t)}\right)\right\}$ 是单调递减序列。又因为 $J\left(W^{(t)},S^{(t)}\right)\ge0$，所以迭代序列 $\left\{J\left(W^{(t)},S^{(t)}\right)\right\}$ 有下界。根据单调有界定理可知,有下界的单调递减序列必有极限,所以 $\left\{J\left(W^{(t)},S^{(t)}\right)\right\}$ 有极限,这说明算法 1 收敛。

3 实验结果与分析

3.1 实验数据描述和实验设置

本文以全连接的热核权重图为基础,对应的图初始关联矩阵为 S,其中 $S_{ij}^{0}=e^{-\frac{\|x_i-x_j\|^2}{2\sigma^2}}$ 表示 x_i 和 x_j 的关联程度。本文选取 9 个数据集进行实验,它们分别是 iris、leaf4、sonar、chessboard、wine、glass、heart、Balance_scale 和 breast_cancer,均来自于 UCI 数据集(http://archive.ics.uci.edu/ml/datasets.html;),其信息如表 1 所示。实验中,对比方法共 3 种,分别是 Nuct[7]、DCD[16]和本文的 GoDCD 方法。对比的指标是用聚类纯度（cluster purity）[16]来表达聚类的精确度。聚类纯度定义为

$$CP=\frac{1}{n}\sum_{k=1}^{r}\max_{1\le l\le r}\{n_{kl}\}\tag{5}$$

表 1 实验使用的数据集及其信息

数据	样本	维数	类别
iris	150	4	3
leaf4	40	14	4
sonar	208	60	2
chessboard	100	2	2
wine	178	13	3
glass	214	9	6
heart	270	13	2
Balance_scale	625	4	3
breast_cancer	682	9	2

其中：n 为数据集中的样本总数，n_{kl} 为算法聚类后属于第 k 类，但在原数据中属于第 l 类的样本数（即：n_{kl} is the number of data samples in the cluster k that belong to ground-truth class l）[16]。显然，$0\le CP\le1$，其值越大则表明聚类精确度越高。

3.2 在聚类精确度上的对比

表 2 实验中聚类纯度对比

	Nuct	dcd	GoDCD
iris	0.8933	0.72	**0.9067**
Sonar	0.5337	0.5433	**0.5721**
chessboard	0.54	**0.57**	**0.57**
Wine	0.6742	0.6798	**0.6966**
glass	0.5047	0.5467	**0.5606**
Heart	0.5556	**0.6074**	**0.6074**
leaf4	0.8	0.475	**0.525**
Balance_scale	0.7832	0.7104	**0.7392**
breast_cancer	0.9589	0.9399	**0.9707**

聚类纯度是表示聚类标签值与真实标签值的相合程度。为了验证 GoDCD 的有效性,在表 2 中列出了聚类纯度对比结果,从中可以看出：

a)DCD 在 iris、sonar、wine、sonar、chessboard 和 heart 上的聚类纯度要明显高于 Ncut 方法。其原因是,相比于 Ncut 方法,DCD 方法不但考虑和利用了数据的图结构,而且还利用低秩双随机分解来增强聚类效果。

b)除了在 chessboard 和 heart 数据集外,GoDCD 的聚类纯度明显高于 DCD。特别地,在 iris 数据集上,GoDCD 方法要比 DCD 方法高出 20%左右。其原因是 DCD 仅考虑在初始图构建上聚类

最优,而 GoDCD 同时考虑了图构建和聚类二者联合最优。

c)在 *leaf4* 数据集合上,GoDCD 和原始 DCD 方法都没有 Ncut 高。其原因可能是双随机分解方法不适合该数据集。

3.3　模型参数对聚类纯度的影响

3.3.1 参数 α 的影响

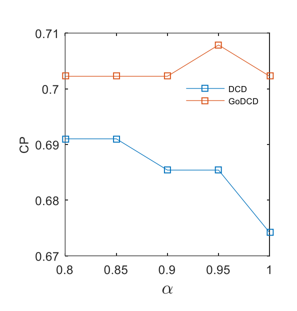


图 1　数据集 *wine* 上不同 α 参值对应的聚类纯度

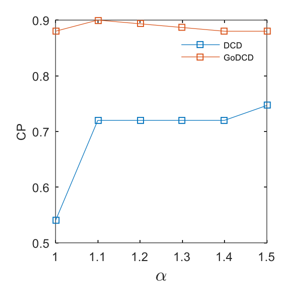


图 2　数据集 iris 上不同 α 参数数值对应的聚类纯度

若聚类纯度相对于参数 α 的不同取值上下浮动较小,则表明算法相对于 α 较稳定。图 1 和图 2 分别给出了 DCD 与 GoDCD 在 wine 和 iris 数据集上，使用不同 α 值所对应的聚类纯度。从图 1 可看出,相比于 DCD，GoDCD 对不同 α 值时的精度波动更小,这说明 GoDCD 比 DCD 更稳定。

3.3.2 参数 λ 的影响

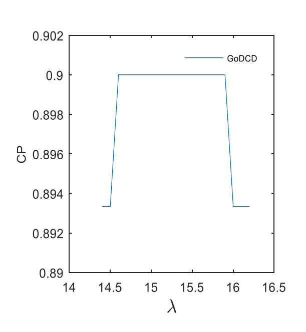


图 3　数据集 *iris* 上不同 λ 值对应的聚类纯度

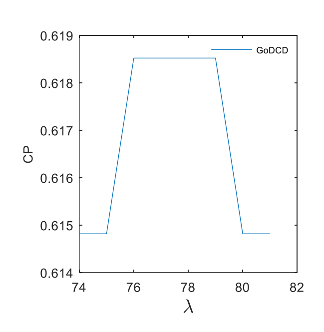


图 4　数据集 *heart* 上不同 λ 值对应的聚类纯度

参数 λ 是 GoDCD 相对于 DCD 增加的一个模型参数。若 λ 太大,则会出现聚类严重依赖初始关联矩阵；若 λ 太小,则会出现更新后的关联矩阵远离数据的原结构。然而,λ 的选择属于模型选择问题,如何选择最优尚无可靠理论。图 2 显示 GoDCD 在 heart 和 iris 数据集上,使用 λ 的不同值所对应的聚类纯度。从图 3 和 4 中可看出,当 λ 在一定的范围内变化时,聚类纯度随着 λ 的变化表现较平缓,这部分地说明 GoDCD 相对于参数 λ 是较稳定的。其原因之一是：即使初始关联矩阵选取得不太好,但由于其被优化,所以 GoDCD 减轻了对初始关联矩阵的依赖程度。

4　结束语

为提高聚类效果,本文提出了图优化的双随机矩阵分解聚类方法 GoDCD,这推广了原始的 DCD 聚类方法。在 GoDCD 中，图优化和低秩双随机分解聚类被集成在同一个目标函数中,其作用是使图构建和聚类达到联合最优,从而减轻了后续聚类对初始图构建质量的依赖程度。在部分 UCI 数据集上的聚类实验结果表明,在大多数情况下,GoDCD 方法比 DCD 方法具有更高的聚类精确度和更好的稳定性。

本文中的 GoDCD 方法仅用于无监督聚类问题,然而有效的半监督辅助信息将有助于实现更精确聚类。因此，如何将 GoDCD 扩展到半监督聚类情形是下一个值得探讨的问题。

参考文献：

[1] Hornik K, Feinerer I, Kober M, et al. Spherical K-means clustering [J]. Journal of Statistical Software, 2012, 50 (10): 1-22.

[2] Park H S, Jun C H. A simple and fast algorithm for K-medoids clustering [J]. Expert Systems with Applications, 2009, 36 (2): 3336-3341.

[3] Wang W Y, Wang C X, Wang J. Research on Hybrid parallel programming technique based on cmp multi-cure cluster [J]. Computer Science, 2014, 41 (2): 19-22.

[4] Festing M, Royer S, Steffen C. Do clusters help firms to realise competitive advantage?A resource-based analysis of the mechanical watch cluster in Glashütte//Germany [J]. Zeitschrift Für Management, 2010, 5 (2): 165-185.

[5] Pan D, Zhao L. Uncertain data cluster based on DBSCAN [C]// Proc of International Conference on Multimedia Technology. 2011: 3781-3784.

[6] Samsonova E V, Kok J N, Ijzerman A P. TreeSOM: cluster analysis in the self-organizing map [J]. Neural Networks the Official Journal of the International Neural Network Society, 2006, 19 (6-7): 935.

[7] Lagrange M, Martins L G, Murdoch J, et al. Normalized cuts for predominant melodic source separation [J]. IEEE Trans on Audio Speech & Language Processing, 2008, 16 (2): 278-290.

[8] 李建元，周脚根，关信红，等. 谱图聚类算法研究进展 [J]. 智能系统学报, 2011, 06 (5): 405-414.

[9] Luxburg U V. A tutorial on spectral clustering [J]. Statistics and Computing, 2007, 17 (4): 395-416.

[10] Arya S, Malamatos T, Mount D M. Space-time tradeoffs for approximate nearest neighbor searching [J]. Journal of the Acm, 2009, 57 (1): 1-54.

[11] He X, Niyogi P. Locality preserving projections [J]. Advances in Neural Information Processing Systems, 2003, 16 (1): 186-197.

[12] Belkin, Mikhail, Niyogi, et al. Laplacian Eigenmaps for dimensionality reduction and data representation [J]. Neural Computation, 2014, 15 (6): 1373-1396.

[13] Maier M, Luxburg U V, Heiny M. of graph construction on graph-based clustering measures [J]. Nips, 2009 (2009): 1025-1032.

[14] Hofmann T. Probabilistic latent semantic indexing [C]// Proc of International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1999: 50-57.

[15] Lee D D, Seung H S. Algorithms for non-negative matrix factorization [C]// Proc of International Conference on Neural Information Processing Systems. MIT Press, 2000: 535-541.

[16] Yang Z, Corander J, Oja E. Low-rank doubly stochastic matrix decomposition for cluster analysis [J]. Journal of Machine Learning Research, 2016, 17 (1): 6454-6478.

[17] Zhang L, Chen S, Qiao L. Graph optimization for dimensionality reduction with sparsity constraints [J]. Pattern Recognition, 2012, 45 (3): 1205-1210.

[18] Zhang L, Qiao L, Chen S. Graph-optimized locality preserving projections [J]. Pattern Recognition, 2010, 43 (6): 1993-2002.

[19] Ding C, Li T, Jordan M I. Nonnegative Matrix Factorization for Combinatorial Optimization: Spectral Clustering, Graph Matching, and Clique Finding [C]// Proc of the 8th IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2008: 183-192.

[20] Arora R, Gupta M R, Kapila A, et al. Similarity-based Clustering by Left-Stochastic Matrix Factorization [J]. Journal of Machine Learning Research, 2013, 14 (1): 1715-1746.